Question 1: Assignment Summary

Briefly describe the "Clustering of Countries" assignment you just completed within 200-300 words. Mention the problem statement and the solution methodology that you followed to arrive at the final list of countries. Explain your main choices briefly(what EDA you performed, which type of Clustering produced a better result, and so on)

Note: You don't have to include any images, equations, or graphs for this question. Just text should be enough.

Answer :

The task was to identify the list of country that requires aid from Help international, initially, we did not have any labels to classify, hence it turned out to be an unsupervised-clustering problem.

The dataset contained 9 attributes like child mortality, gdpp, income, etc, There were no duplicated rows and no missing values found in it. But we had to convert some variables from percentage to absolute values as part of the cleaning process. And no type casting was required in the dataset. We used standard scaling, (not normalization) for scaling since we had the risk of outlier influence while scaling. As we found that many of the attributes are highly correlated, we choose not to directly use the features in the model instead we put the principal component analysis (PCA) in action to explain the variance of the dataset. The first 3 components of PCA were able to explain 87% of the variance in the dataset. After the PCA we found some statistical outliers present in the arrived data frame, we choose to cap them between the 5th percentile and 99th percentile, but not to remove any records because we wanted all the listed countries to be clustered and at the same time the outliers should not skew the clusters being formed. And then we run the Hopkins statistics to check the dataset is good enough for clustering, and we found the Hopkins score = 0.83, which is good.

We tried both KMeans clustering and hierarchical clustering to check which approach gives the better result. In Kmeans, we employed the Elbow curve and silhouette analysis to find out the optimum value of K. we found that k = 4 and k = 5 could be ideal in this problem. And in hierarchical clustering, we used both single and complete linkages and found that the complete linkage performs better.

After building and visualizing the models, we came to know no model is performing exceptionally well as all of them show some high inter-cluster distances. Anyway, the model built on Kmeans with K = 5 was found to be the better one among all the models.

After finalizing the model, we got two clusters that are worth considering, cluster 0 and cluster 4 which were found to be the needing aid from Help-International. These clusters are high in child mortality, low in income low in gdpp, and high in inflation. initially, there were 69 countries on the list and we deduced the list based on the mean (average) values of key drivers like - child mortality, and the income of the country, then we short-listed 23 countries that are in desperate need of aid from the world, and hence from the Help-International.

Question 2: Clustering

a) Compare and contrast K-means Clustering and Hierarchical Clustering.

Answer :

K Means Clustering

k-means, using a pre-specified number of clusters, the method assigns records to each cluster to find the mutually exclusive cluster of spherical shape based on distance. K Means clustering needed advanced knowledge of K i.e. no. of clusters one wants to divide your data. One can use median or mean as a cluster center to represent each cluster. Methods used are normally less computationally intensive and are suited to very large datasets.

Hierarchical Clustering

In hierarchical clustering one can stop at any number of clusters, one finds appropriate by interpreting the dendrogram. Hierarchical methods can be either divisive or agglomerative. Agglomerative methods begin with 'n' clusters and sequentially combine similar clusters until only one cluster is obtained. Divisive methods work in the opposite direction, beginning with one cluster that includes all the records. And Hierarchical methods are especially useful when the target is to arrange the clusters into a natural hierarchy. In Hierarchical Clustering, results are reproducible in Hierarchical clustering

b) Briefly explain the steps of the K-means clustering algorithm.

Answer :

1. Select initial centroids based on the input regarding the number of centroids that should be given by the user.
2. Assign the data points to the closest centroid
3. Recalculate the centroid for each cluster and assign the data objects again
4. Follow the same procedure until convergence. Convergence is achieved when there is no more assignment of data objects from one cluster to another, or when there is no change in the centroid of clusters.

c) How is the value of 'k' chosen in K-means clustering? Explain both the statistical as well as the business aspect of it.

Answer :

There are two methods :

1. The Elbow Method
2. The Silhouette Method

The Elbow Method

In cluster analysis, the elbow method is a heuristic used in determining the number of clusters in a data set. The method consists of plotting the explained variation as a function of the number of clusters and picking the elbow of the curve as the number of clusters to use. In clustering, this means one should choose a number of clusters so that adding another cluster doesn't give much better modeling of the data. The intuition is that increasing the number of clusters will naturally improve the fit (explain more of the variation) since there are more parameters (more clusters) to use, but that at some point this is over-fitting, and the elbow reflects this

The Silhouette Method

In the Silhouette algorithm, we assume that the data has already been clustered into k clusters by a clustering technique (Typically the K-Means Clustering technique).
Then for each data point, we define the following:-

(i) -The cluster assigned to the ith data point
(i) – The number of data points in the cluster assigned to the ith data point
(i) – It gives a measure of how well assigned the ith data point is to its cluster

$a(i) = \frac{1}{|C(i)|-1} \sum_{C(i), i \neq j} d(i,j)$

b(i) – It is defined as the average dissimilarity to the closest cluster which is not it's cluster

$b(i) = min_{i \neq j} (\frac{1}{|C(j)|} \sum_{j \in C(j)} d(i,j))$

The silhouette coefficient s(i) is given by:-

$s(i) = \frac{b(i)-a(i)}{max(a(i),b(i))}$

NOTE: The silhouette Method is used in combination with the Elbow Method for a more confident decision.

Business Aspect of Choosing the Number of K's

Considering both the above method to arrive at K, we can choose k which is most understandable to business and practically used for future activity so that it can be tracked.

Suppose, as per the above calculation, we May 15 customer classifications, but that is not practically interpretable or implemented in the business, hence the above methods with business requirements can together give you a number which could be considered as optimum k.

d) **Explain the necessity for scaling/standardization before performing Clustering.**

Answer :

For Clustering, we need a way to compute the distance between pairs of data points. Data points that are close to each other will more likely belong to the same cluster.

Let's take k-means as the clustering method. Each data point is represented as a point in space. If the dataset has 1 feature, the space we use is 1-dimensional; if the dataset has 2 features, the space we used is 2-dimensional; etc. The distance between a pair of data points is computed as the Euclid distance between the 2 corresponding points.

The reason we normalize the data is to make sure all dimensions are treated equally. In other words, we want each column to contribute the same impact on the distance. Note that normalization is done on each column separately (rather than on each row).

Also, note that there are many ways to normalize, you may standardize (de-mean and then divide by the standard deviation) or min-max scale (scale all columns so that in each column, the minimum value is 0 and the maximum value is 1).

e) Explain the different linkages used in Hierarchical Clustering.

Answer :

Single Linkage:

The distance between 2 clusters is defined as the shortest distance between points in the two clusters

Complete Linkage:

The distance between 2 clusters is defined as the maximum distance between any 2 points in the clusters

Average Linkage:

The distance between 2 clusters is defined as the average distance between every point of one cluster to every other point of the other cluster.

Note :

You have to decide what type of linkage should be used by looking at the data. One convenient way to decide is to look at how the dendrogram looks. Usually, a single linkage-type will produce dendrograms that are not structured properly, whereas a complete or average linkage will produce clusters that have a proper tree-like structure.